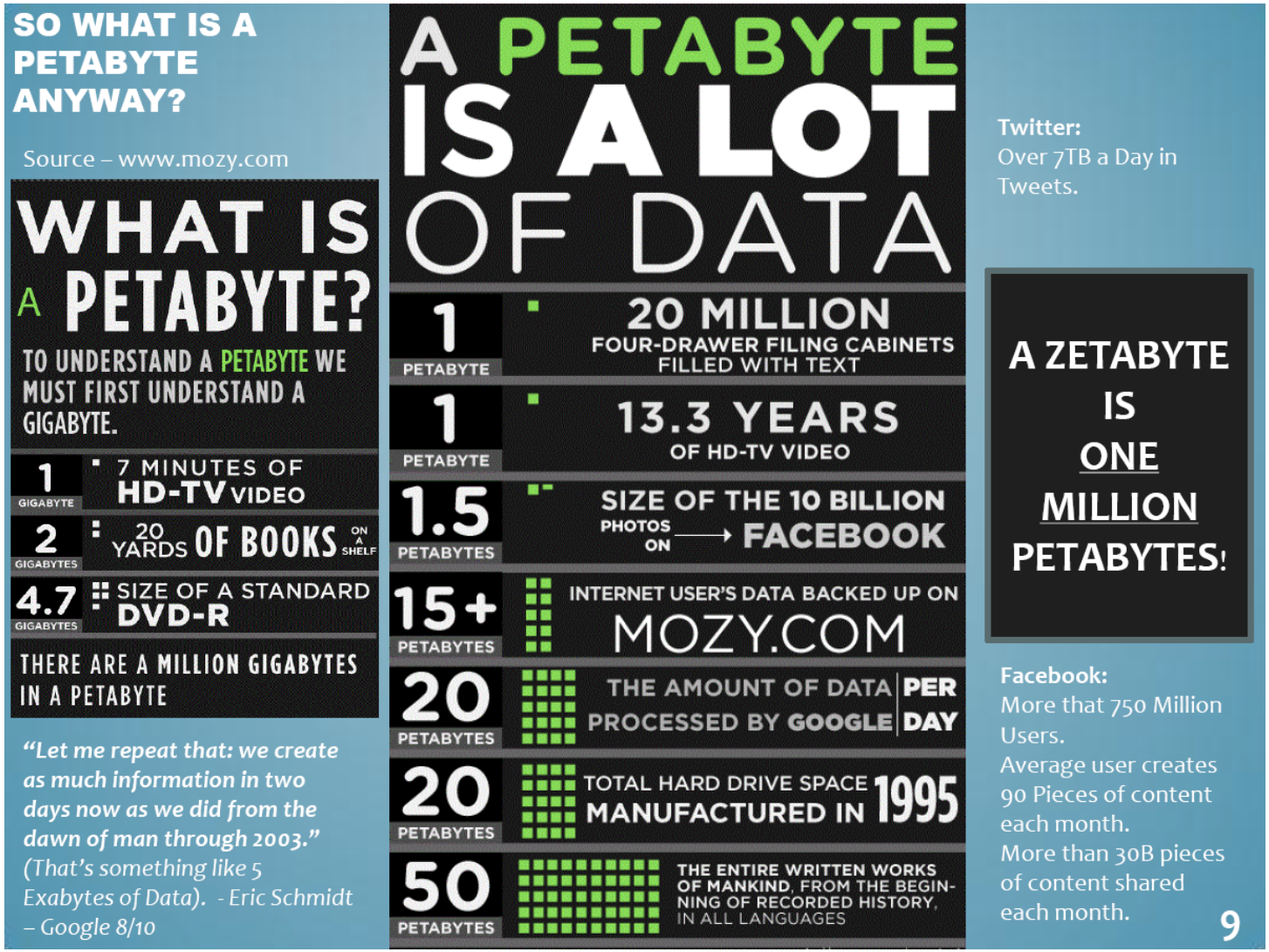
Future lessons from large-scale biological data management

Paul Flicek
Vertebrate Genomics, European Bioinformatics Institute
European Molecular Biology Laboratory
Wellcome Trust Sanger Institute

EMBL-EBI

We have been living through a revolution.

Revolution is driven by data



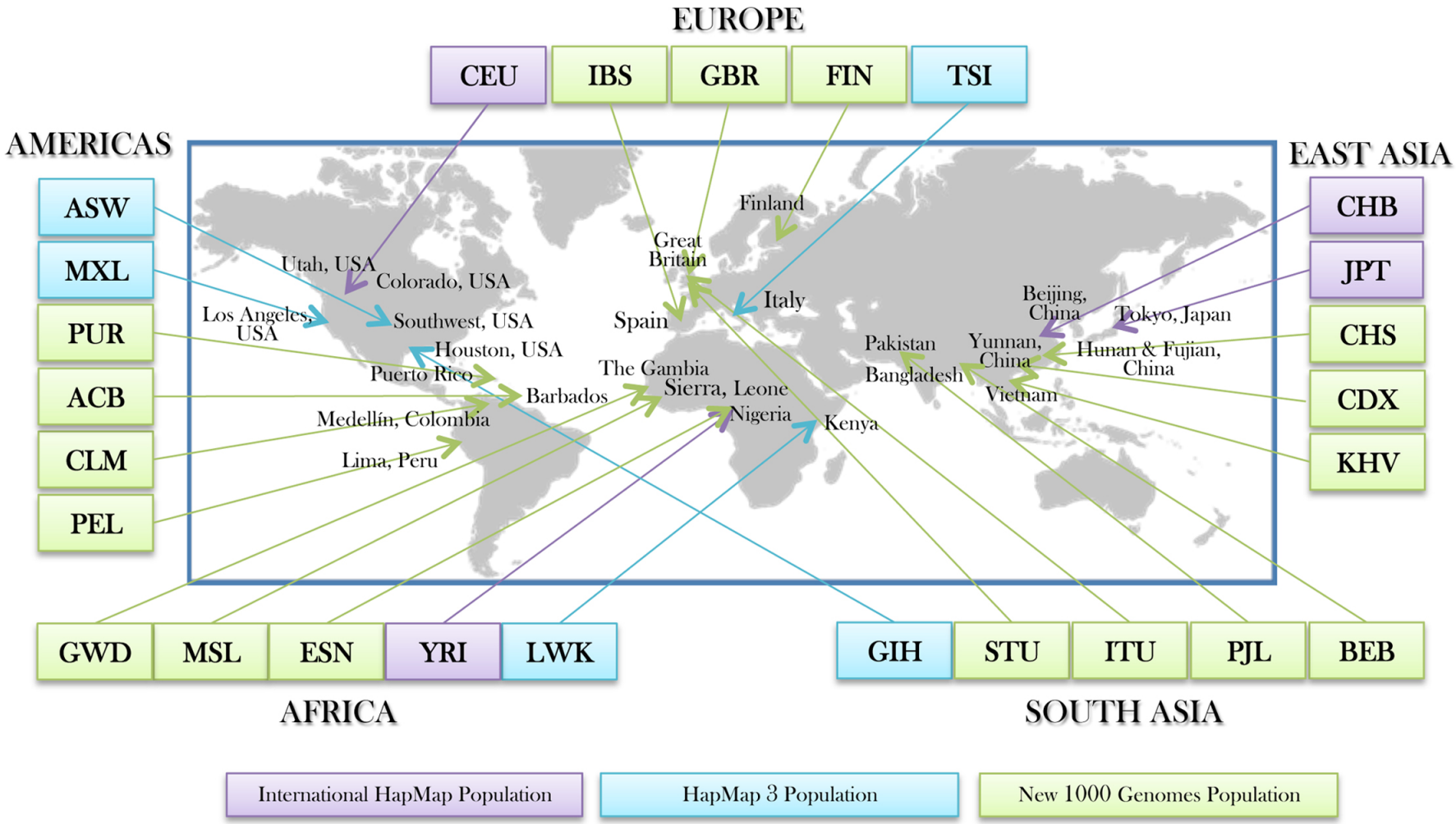
Source: Semantic community

A data driven experiment: The 1000 Genomes Project primary goals

To provide a deep characterization of human genome variation to provide a baseline for investigating the relationship between genotype and phenotype.

- To identify effectively all variation
  - At 1% MAF or higher genome wide
  - At 0.1% to 0.5% MAF in the exonic regions
- Structural variation as well as SNVs
- Provide a haplotype structure for the human genome
- Develop analysis methods, tools and reagents which can be transferred to other projects

Which samples?



Basic strategy

- Collect shotgun sequencing reads
- Random Fragments of the whole genome or exome
- Map the reads to the reference genome
  - Possible problems with repetitive regions of the genome
  - Possible problems with misalignments
- Detect variation based on the alignment of the reads from all samples
  - Statistical issues allowing for errors in sampling

Seven years ago little of this could be done at scale

The growth of the project

- Pilot (2008-2010, published Nature Oct. 2010):
  - Deep sequence for two trios (CEU and YRI)
  - Low coverage (~2x) of 180 individuals in 3 populations
  - Capture of 1000 genes in ~700 individuals

- Phase 1 (2010-2012, published Nature Nov. 2012)
  - 1092 individuals with ~3x low-coverage, 1040 with matched exome sequence
  - OMNI 2.5M genotyping

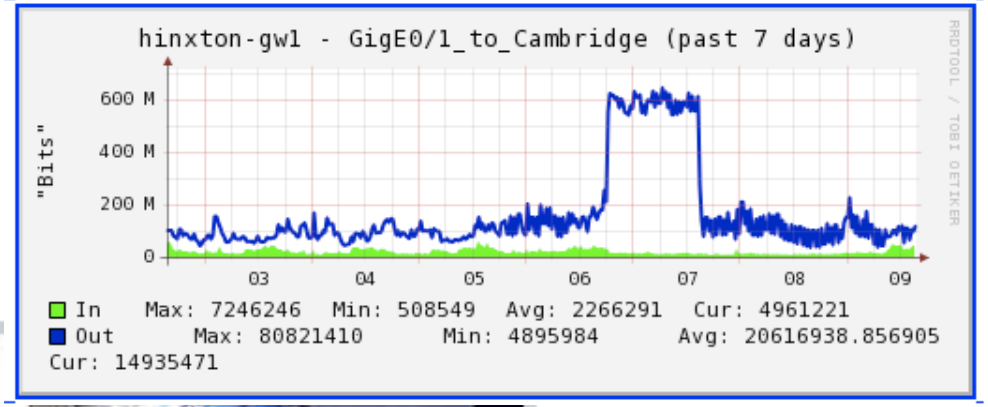
- Phase 2+3 (2012-2014, publish final Paper TBD)
  - 2535 samples with low coverage and exome sequence data
  - High coverage Complete Genomics data for 427 samples.

- Final project represents 25 times more data than the original plan. With 2.5 times more samples and more populations

Managing the 1000 Genomes data

What it felt like in April 2008

First major data transfer



Today

Infrastructures are critical…

But we only notice them when they go wrong

Informatics is Infrastructure:

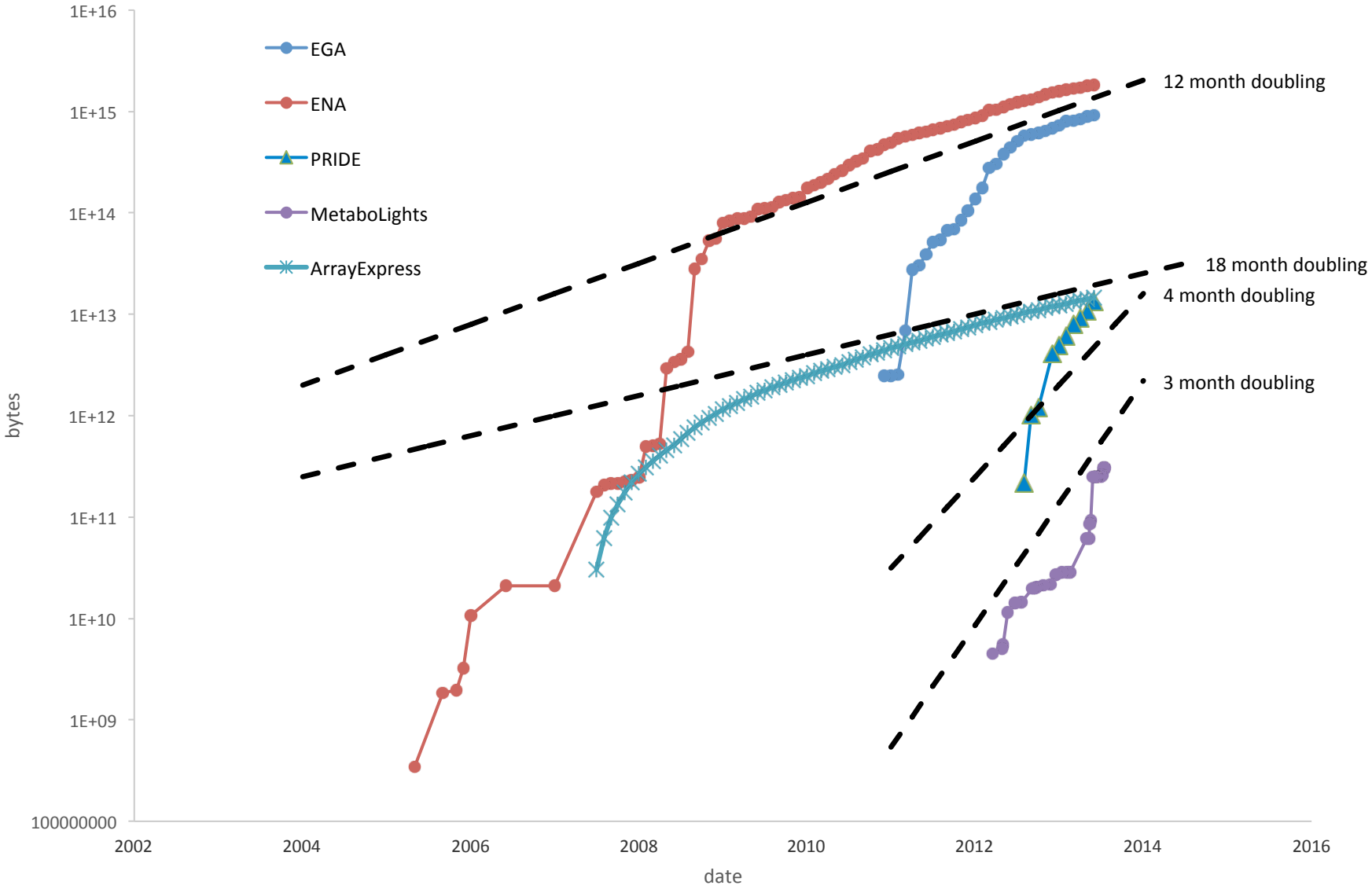
Network transfer protocols
Data Compression
Standards
Archives

1000 Genomes Project Size

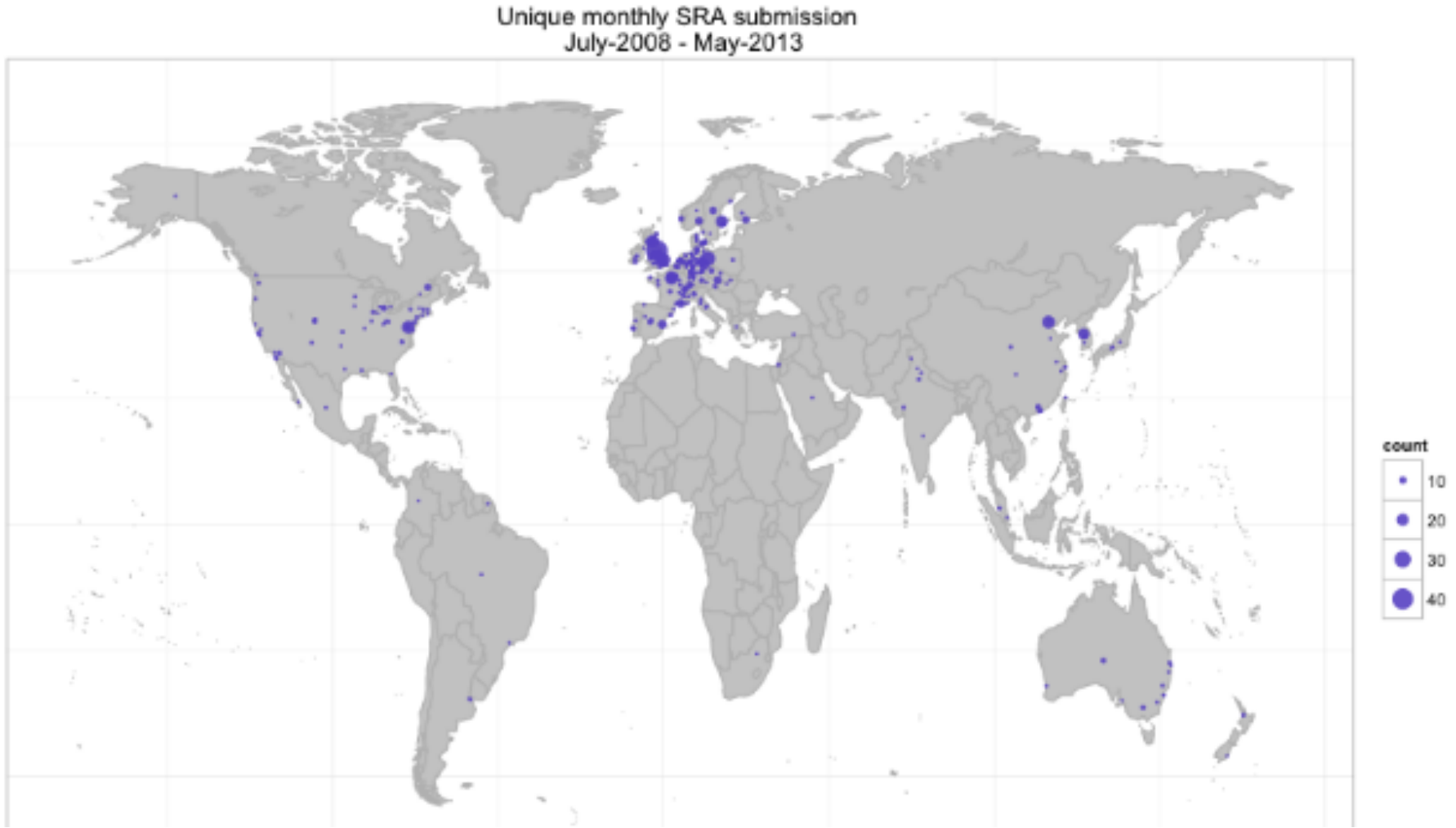
- There are 4461406 files on the ftp site
- There are 580T of data on the ftp site
- There are 26 populations
- There are 2854 samples
- There are 79072 gigabases of low coverage sequence
- 28753 x coverage in low coverage
- There are 35607 gigabases of exome sequence

There are currently 1,196,200 GB of sequence in the ENA in total (was 235 GB at the start of the project).

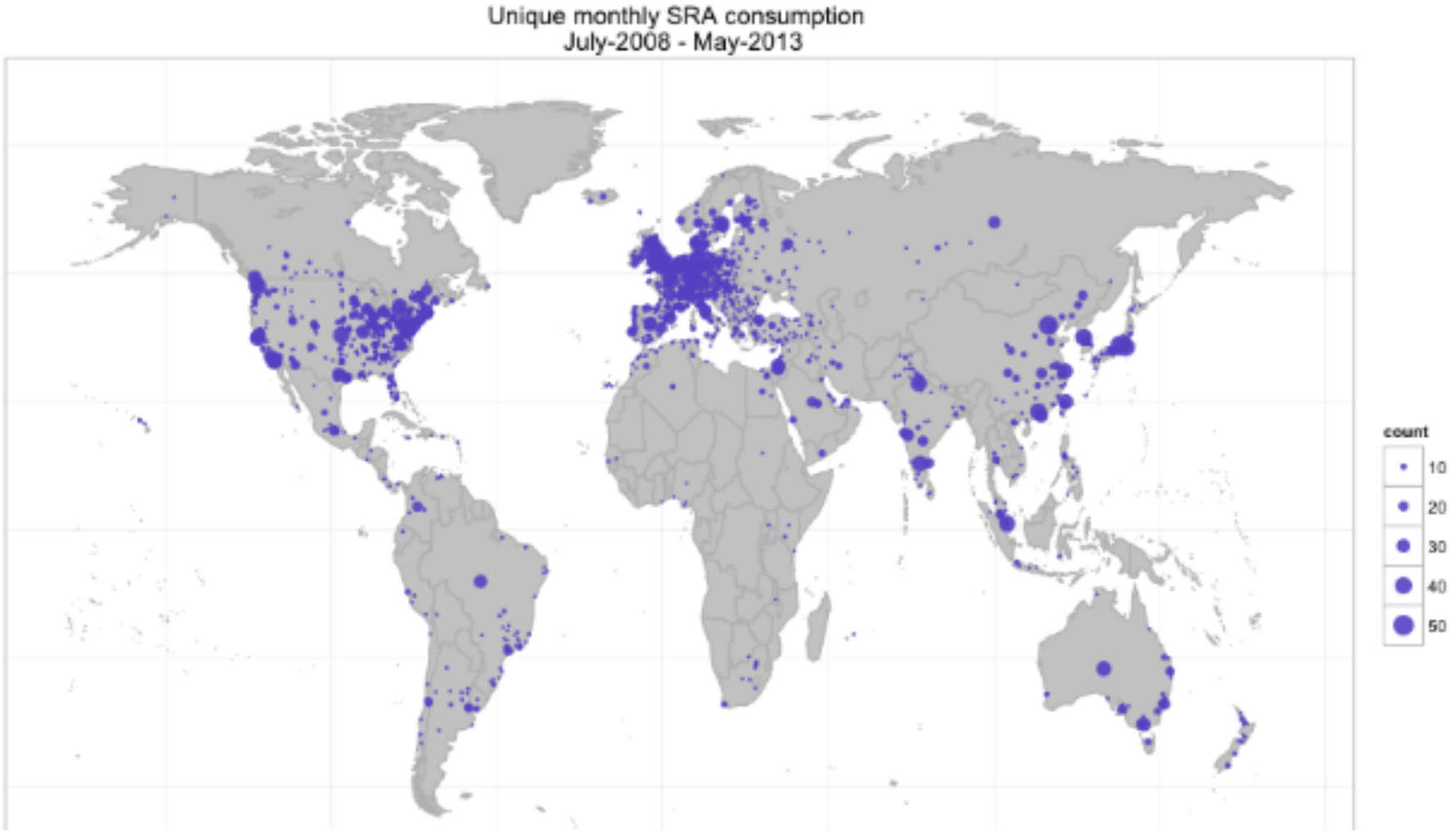
Data growth


1E+16
1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
1E+09
100000000
bytes
2002
2004
2006
2008
2010
2012
2014
2016
date
EGA
ENA
PRIDE
MetaboLights
ArrayExpress
12 month doubling
18 month doubling
4 month doubling
3 month doubling

Distributed production: sequence data submission



Distributed consumption: sequence data access



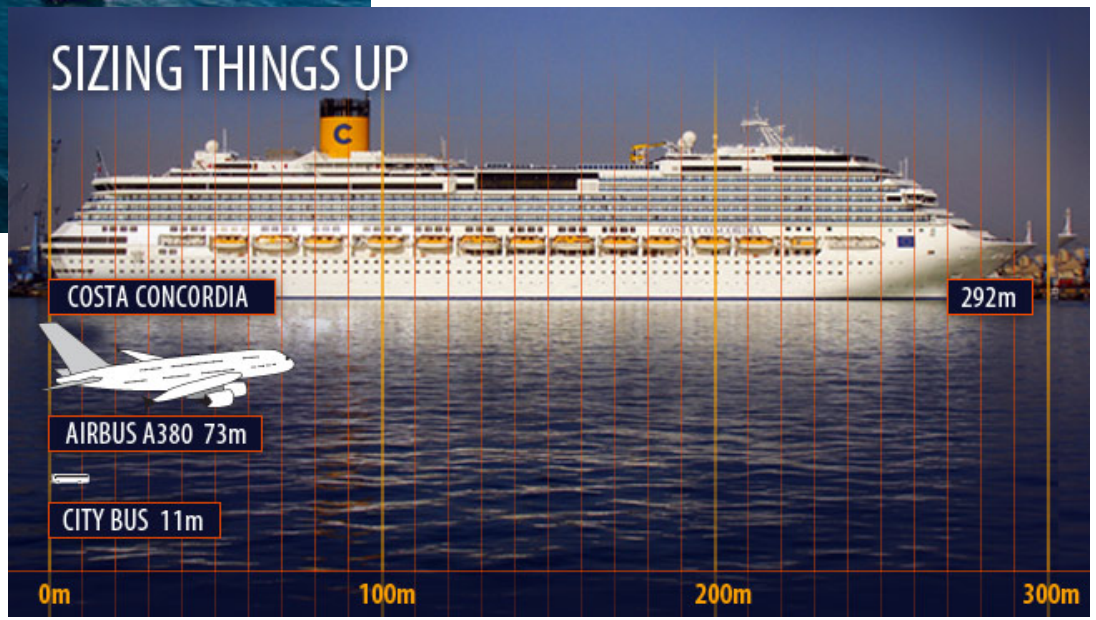
In how many ways can you say “female”?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

Big problems need solutions

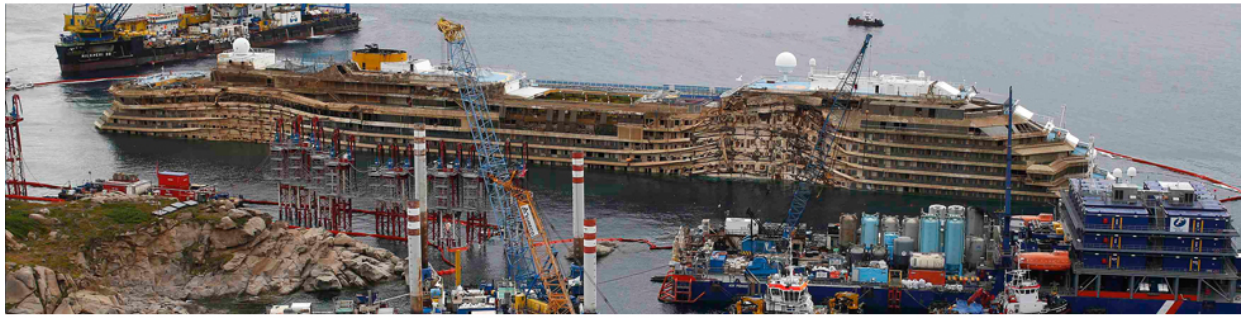
Source: Guardian.co.uk



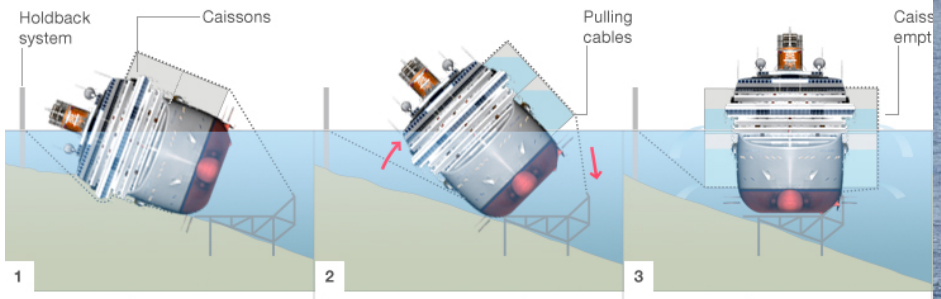
Graphic: Matt Pike Source: News Limited

Solutions are often possible

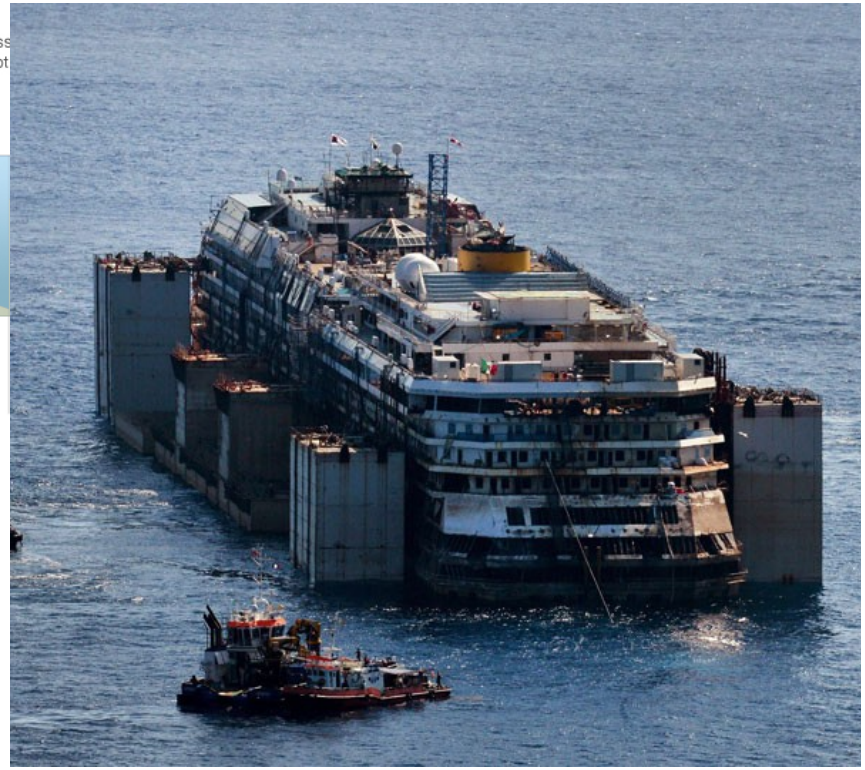
Costa Concordia, September 2013



Salvage operation



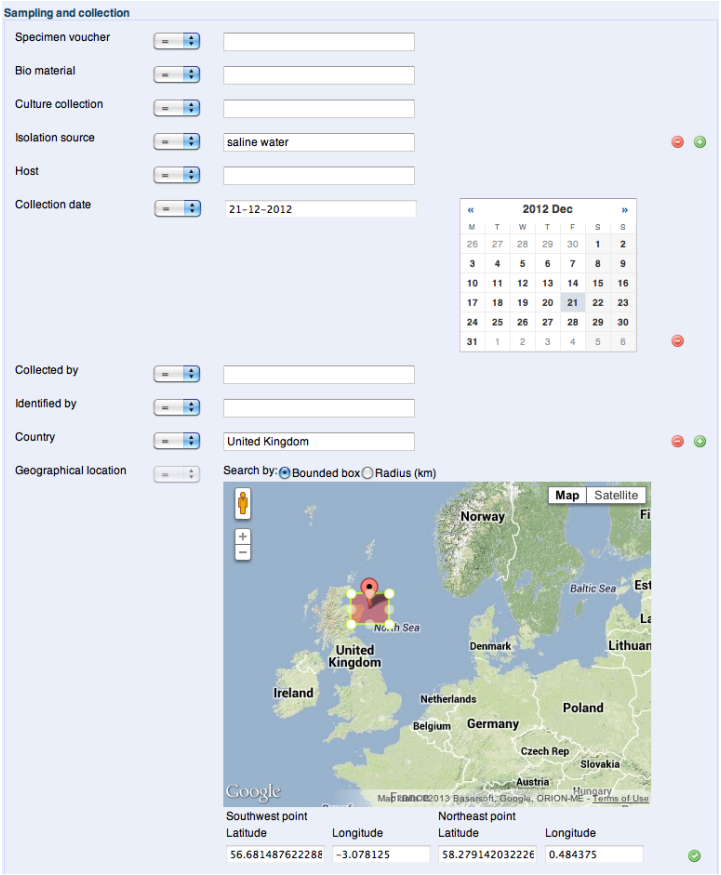
Source: Titan/Micoperi. Image: Getty



Informatics is Infrastructure:

Network transfer
Data Compression
Standards
Archives

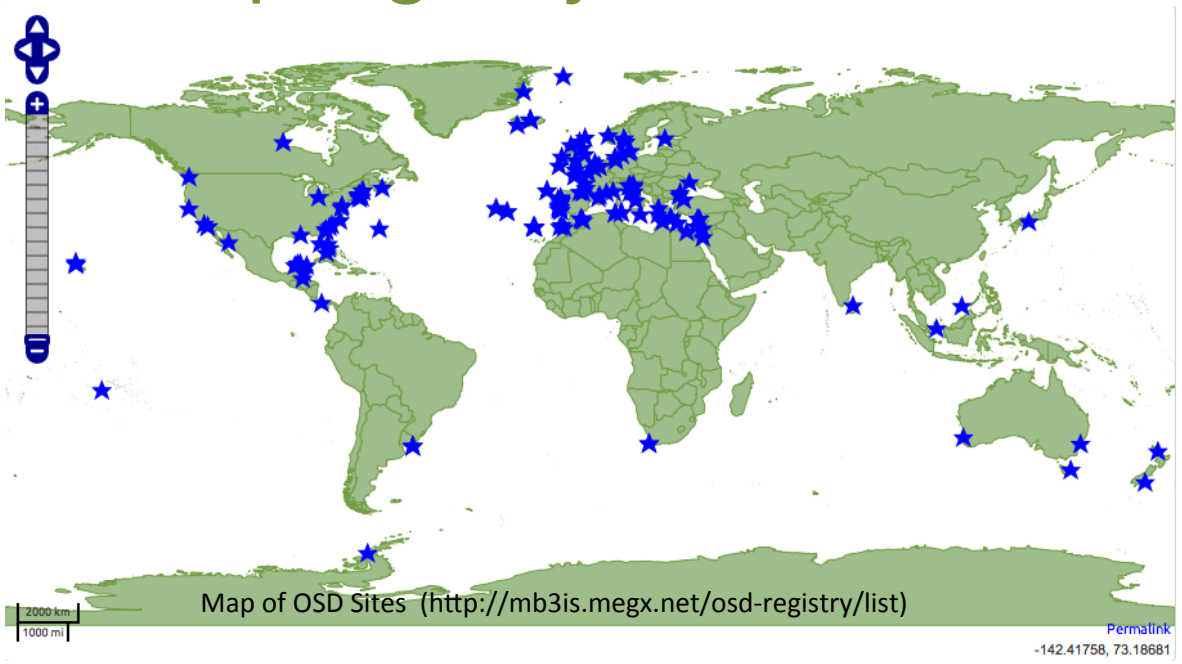
Standards-compliant data are more discoverable



search terms specified for fields '**isolation source**', '**collection date**' and '**geographical location**'

Ocean Sampling Day

http://www.microb3.eu/osd



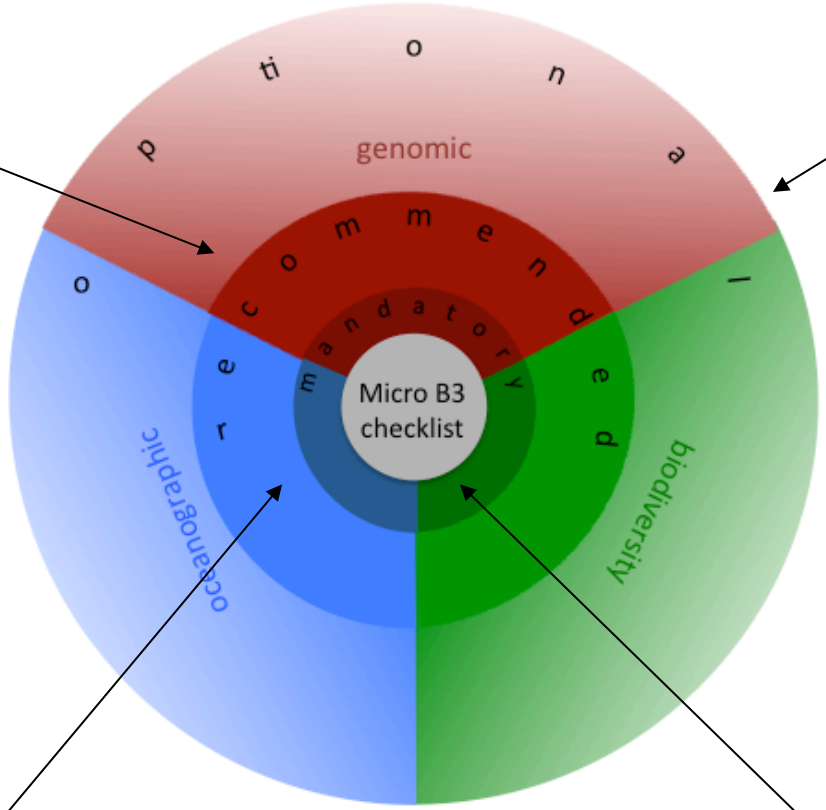
Map of OSD Sites (http://mb3is.megx.net/osd-registry/list)

- 160 marine stations ran simultaneous standardised sampling on the 21st of June 2014
- A snapshot into status of world's oceans and seas generating a reference dataset that will provide insight into marine microbial diversity and function in the marine environment
- Cross-discipline geographically informed contextual data reporting

Reporting standards for OSD

recommended
(for rich environmental context and data reproducibility)

optional
(for comprehensive description of a sample and its environment)



mandatory
(for meaningful analysis and correct data processing)

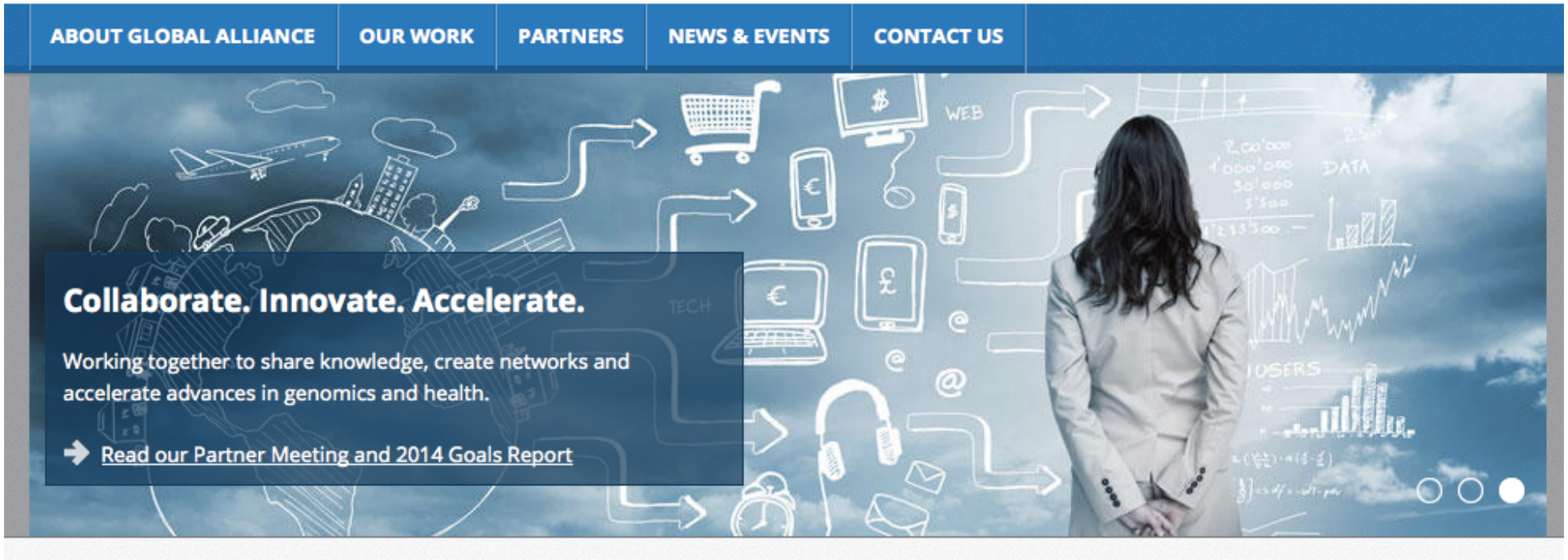
reporting minimum
(OSD Sites report, includes interoperability minimum)

Informatics is Infrastructure:

The Future

The Future

- Infrastructure and standards allow for us to imagine and realise bigger projects in the future
- A connected web of domain specific efforts, general projects, technology and innovation will drive this infrastructure
  - The Global Alliance for Genomics and Health
  - ELIXIR
  - Secure cloud-based computing
  - Software

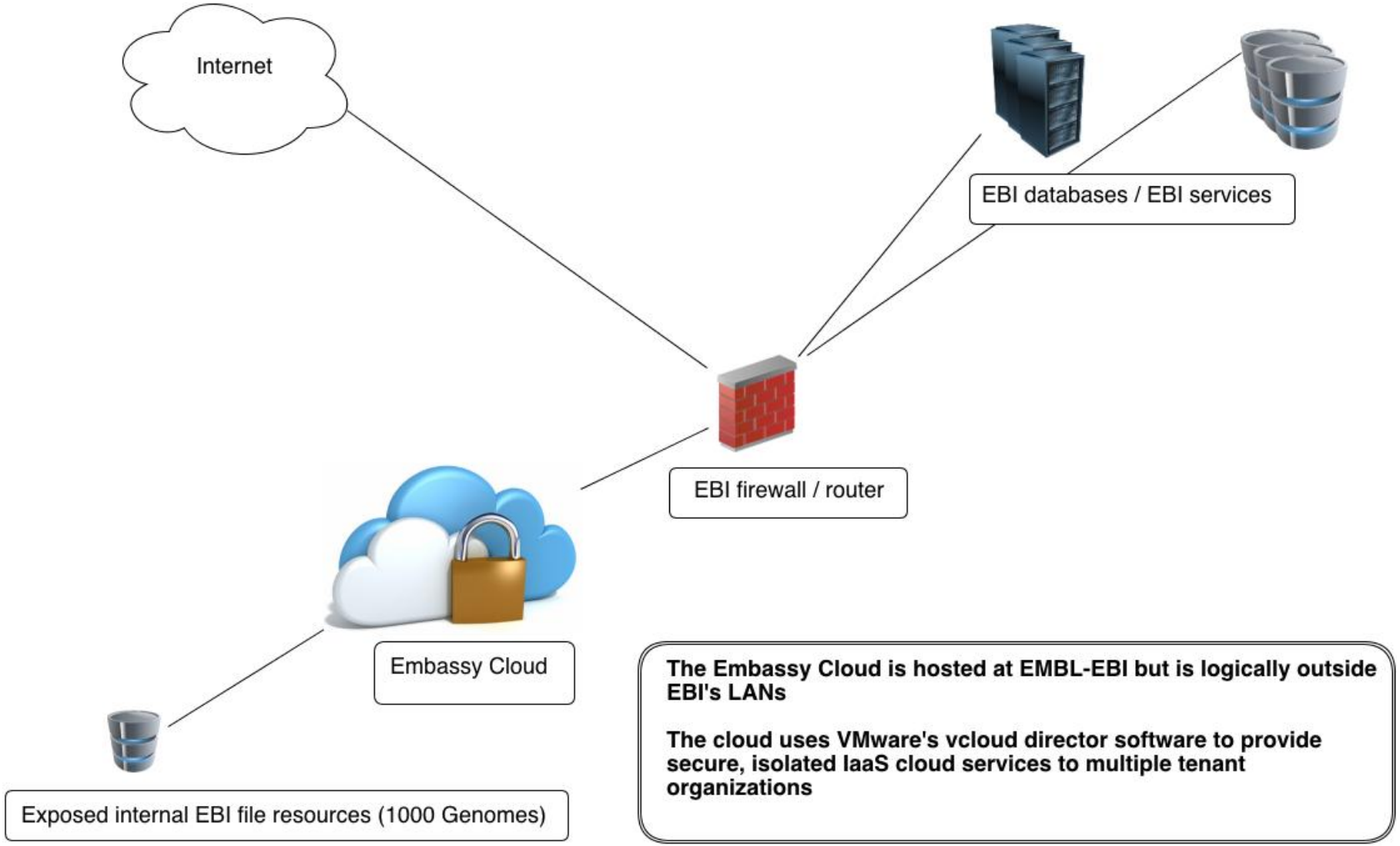


- Technology, standards and protocols for federated sharing and analysis of human genomic and health data
- Applicable to many problem is data management

Building capacity in Europe

- ELIXIR: a sustainable infrastructure for biological information in Europe.
- Supporting life science research and its translation to:
  - medicine
  - agriculture
  - the environment
  - the bioindustries
  - society.
- Supported by UK LFCF

EMBL-EBI Embassy Cloud



The Embassy Cloud is hosted at EMBL-EBI but is logically outside EBI's LANs

The cloud uses VMware's vcloud director software to provide secure, isolated IaaS cloud services to multiple tenant organizations

Infrastructure enables discovery

Necessary (if conceptually unexciting) data management

Interesting, ground breaking ideas

Acknowledgements

People

- Guy Cochrane and the ENA Team
- Laura Clarke and the 1000 Genomes Project
- Andy Cafferkey and EBI's Cloud Team
- Justin Paschall and the EBI Variation Archive Team
- Fiona Cunningham and the entire Ensembl Team

Funding

wellcome trust

EMBL

National Human Genome Research Institute

BBSRC bioscience for the future

European Commission Framework Programme 7

SEVENTH FRAMEWORK PROGRAMME

European Bioinformatics Institute

EMBL